Preventing Over-Fitting of Cross-Validation with Kernel Stability

Yong Liu and Shizhong Liao*

School of Computer Science and Technology
Tianjin University, Tianjin, China
szliao@tju.edu.cn

Abstract. Kernel selection is critical to kernel methods. Cross-validation (CV) is a widely accepted kernel selection method. However, the CV based estimates generally exhibit a relatively high variance and are therefore prone to over-fitting. In order to prevent the high variance, we first propose a novel version of stability, called kernel stability. This stability quantifies the perturbation of the kernel matrix with respect to the changes in the training set. Then we establish the connection between the kernel stability and variance of CV. By restricting the derived upper bound of the variance, we present a kernel selection criterion, which can prevent the high variance of CV and hence guarantee good generalization performance. Furthermore, we derive a closed form for the estimate of the kernel stability, making the criterion based on the kernel stability computationally efficient. Theoretical analysis and experimental results demonstrate that our criterion is sound and effective.

1 Introduction

Kernel methods, such as support vector machine (SVM) [36], kernel ridge regression (KRR) [32] and least squares support vector machine (LSSVM) [35], have been widely used in machine learning and data mining. The performance of these algorithms greatly depends on the choice of kernel function, hence kernel selection becomes one of the key issues both in recent research and application of kernel methods [9].

It is common to select the kernel selection for kernel methods based on the generalization error of learning algorithms. However, the generalization error is not directly computable, as the probability distribution generating the data is unknown. Therefore, it is necessary to resort to estimates of the generalization error, either via testing on some data unused for learning (hold-out testing or cross-validation techniques) or via a bound given by theoretical analysis. To derive the theoretical upper bounds of the generalization error, some measures are introduced: such as VC dimension [36], Rademacher complexity [2], regularized risk [33], radius-margin bound [36], compression coefficient [26], Bayesian regularisation [7], influence function [14], local Rademacher complexity [11], and eigenvalues perturbation [23], etc.

* Corresponding author.

T. Calders et al. (Eds.): ECML PKDD 2014, Part II, LNCS 8725, pp. 290–305, 2014.

While there have been many interesting attempts to use the theoretical bounds of generalization error or other techniques to select kernel functions, the most commonly used and widely accepted kernel selection method is still cross-validation. However, the cross-validation based estimates of performance generally exhibit a relatively high variance and are therefore prone to over-fitting [19,27,7,8]. To overcome this limitation, we introduce a notion of kernel stability, which quantifies the perturbation of the kernel matrix when removing an arbitrary example from the training set. We illuminate that the variance of cross-validation for KRR, LSSVM and SVM can be bounded based on the kernel stability. To prevent the high variance of cross-validation, we propose a novel kernel selection criterion by restricting the derived upper bound of the variance. Therefore, the kernel chosen by this criterion can avoid over-fitting of cross-validation. Furthermore, the closed form of the estimate of the kernel stability is derived, making the kernel stability computationally efficient. Experimental results demonstrate that our criterion based on kernel stability is a good choice for kernel selection. To our knowledge, this is the first attempt to use the notion of stability to entirely quantify the variance of cross-validation for kernel selection.

The rest of the paper is organized as follows. Related work and preliminaries are respectively introduced in Section 2 and Section 3. In Section 4, we present the notion of kernel stability, and use this stability to derive the upper bounds of the variance of cross-validation for KRR, LSSVM and SVM. In Section 5, we propose a kernel selection criterion by restricting these bounds. In Section 6, we analyze the performance of our proposed criterion compared with other state-of-the-art kernel selection criteria. Finally, we conclude in the last section.

2 Related Work

Cross-validation has been studied [27,19,3,15] and used in practice for many years. However, analyzing the variance of cross-validation is tricky. Bengio and Grandvalet [3] asserted that there exists no universal unbiased estimator of the variance of cross-validation. Blum et al. [4] showed that the variance of the cross-validation estimate is never larger than that of a single holdout estimate. Kumar et al. [20] generalized the result of [4] considerably, quantifying the variance reduction as a function of the algorithm's stability. Unlike the above work which considers the link between the variance of the cross-validation estimate and that of the single holdout estimate, in this paper we consider bounding the variance of cross-validation for some kernel methods, such as KRR, LSSVM and SVM, based on an appropriately defined notion of stability for kernel selection.

The notion of stability has been studied in various contexts over the past years. Rogers and Wagner [31] presented the definition of weak hypothesis stability. Kearns and Ron [16] defined the weak-error stability in the context of proving sanity check bounds. Kutin and Niyogi [21] defined the uniform stability notion; see also the work of Bousquet and Elisseeff [5]. The notions of mean square stability and the loss stability were introduced by Kumar et al. [20], which are

closely related to the leave-one-out cross-validation. Unfortunately, for most of these notions of stability, proposed to derive the theoretical generalization error bounds, it is difficult to compute their specific values [28]. Thus, these notions of stability are hard to be used in practical kernel selection. To address this issue, we propose a new version of stability, which is defined on a kernel function, are computationally efficient and practical for kernel selection.

3 Preliminaries and Notations

Let $S = \{z_i = (\boldsymbol{x}_i, y_i)\}_{i=1}^n$ be a sample set of size n drawn i.i.d from a fixed, but unknown probability distribution P on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is the input space and $\mathcal{Y}$ is the output space. Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a kernel. The reproducing kernel Hilbert space (RKHS) $\mathcal{H}_K$ associated with K is defined to be the completion of the linear span of the set of functions $\{K(\boldsymbol{x}, \cdot) : \boldsymbol{x} \in \mathcal{X}\}$ with the inner product denoted as $\langle \cdot, \cdot \rangle_K$ satisfying

$$\left\langle \sum_{i=1}^n \alpha_i K(\boldsymbol{x}_i, \cdot), \sum_{i=1}^n \beta_i K(\boldsymbol{x}'_i, \cdot) \right\rangle_K = \sum_{i,j=1}^n \alpha_i \beta_j K(\boldsymbol{x}_i, \boldsymbol{x}'_j).$$

We assume that $|y| \leq M$ for all $y \in \mathcal{Y}$ and $K(x, x) \leq \kappa$ for all $x \in \mathcal{X}$.

The learning algorithms we study here are the regularized algorithms:

$$f_S := \operatorname*{arg\,min}_{f \in \mathcal{H}_K, b \in \mathbb{R}} \left\{ \frac{1}{|S|} \sum_{z \in S} \ell(y_i, f(\boldsymbol{x}_i) + b) + \lambda \|f\|_K^2 \right\},$$

where $\ell(\cdot, \cdot)$ is a loss function, λ is the regularization parameter and $|S|$ is the size of S. KRR, LSSVM, and SVM are the special cases of the regularized algorithms. For KRR,

$$b = 0 \text{ and } \ell(f(\boldsymbol{x}), y) = (y - f(\boldsymbol{x}))^2,$$

for LSSVM

$$\ell(f(\boldsymbol{x}) + b, y) = (y - f(\boldsymbol{x}) - b)^2,$$

and for SVM

$$\ell(f(\boldsymbol{x}) + b, y) = \max\left(0, 1 - y\left(f(\boldsymbol{x}) + b\right)\right).$$

The (empirical) loss of the hypothesis f_S on a set Q is defined as

$$\ell_{f_S}(Q) = \frac{1}{|Q|} \sum_{z \in Q} \ell(f_S(\boldsymbol{x}), y).$$

Let $S_1, \ldots, S_k$ be a random equipartition of S into k parts, called folds, with $|S_i| = \left\lfloor \frac{n}{k} \right\rfloor$. We learn k different hypotheses with $f_{S \setminus S_i}$ being the hypothesis

learned on all of the data except for the ith fold; Let $m = (k-1)k/n$ be the size of the training set for each of these k hypotheses. The k-fold cross-validation hypothesis, $f_{k\text{cv}}$, which picks one of the $\{f_{S\setminus S_i}\}_{i=1}^{k}$ uniformly at random. The (empirical) loss of $f_{k\text{cv}}$ is defined as

$$\ell_{f_{k\text{cv}}}(S) = \frac{1}{k}\sum_{i=1}^{k}\frac{1}{|S_i|}\sum_{z\in S_i}\ell(f_{S\setminus S_i}(\boldsymbol{x}), y).$$

4 Variance Bounds of Cross-Validation

k-fold cross-validation (k-CV) is the most widely accepted method for kernel selection. However, it is known to exhibit a relatively high variance $\operatorname{var}_S\left(\ell_{f_{k\text{cv}}}(S)\right)$,

$$\operatorname*{var}_S\left(\ell_{f_{k\text{cv}}}(S)\right) = \mathop{\mathbb{E}}_{S\sim\mathcal{Z}^n}\left[\ell_{f_{k\text{cv}}}(S) - \mathop{\mathbb{E}}_{S\sim\mathcal{Z}^n}\left[\ell_{f_{k\text{cv}}}(S)\right]\right]^2.$$

Therefore, k-CV is prone to over-fitting [19,27,7,8]. Obviously, $\operatorname{var}_S\left(\ell_{f_{k\text{cv}}}(S)\right)$ is not directly computable, as the probability distribution is unknown. In the next subsection, we will define a new notion of stability to bound $\operatorname{var}_S\left(\ell_{f_{k\text{cv}}}(S)\right)$.

4.1 Kernel Stability

The way of making the definition of kernel stability is to start from the goal: to get bounds on the variance of CV and want these bounds to be tight when the kernel function satisfies the kernel stability.

It is well known that the kernel matrix contains most of the information needed by kernel methods. Therefore, we introduce a new notion of stability to quantify the perturbation of the kernel matrix with respect to the changes in the training set for kernel selection.

To this end, let $T = \{\boldsymbol{x}_i\}_{i=1}^{m}$ and the ith removed set T^i be

$$T^i = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{i-1}, \boldsymbol{x}_{i+1}, \ldots, \boldsymbol{x}_m\}.$$

Denote the kernel matrix $\boldsymbol{K}$ as $[K(\boldsymbol{x}_i, \boldsymbol{x}_j)]_{i,j=1}^{m}$, and let $\boldsymbol{K}^i$ be the $m \times m$ ith removed kernel matrix with

$$\begin{cases} [\boldsymbol{K}^i]_{jk} = K(\boldsymbol{x}_j, \boldsymbol{x}_k) & \text{if } j \neq i \text{ and } k \neq i, \\ [\boldsymbol{K}^i]_{jk} = 0 & \text{if } j = i \text{ or } k = i. \end{cases}$$

One can see that $\boldsymbol{K}^i$ can be considered as the kernel matrix with respect to the removed set T^i.

Definition 1 (Kernel Stability). *A kernel function K is of β kernel stability if the following holds: $\forall x_i \in \mathcal{X}$, $i = 1, \ldots, m$,*

$$\forall i \in \{1, \ldots, m\}, \|\boldsymbol{K} - \boldsymbol{K}^i\|_2 \leq \beta,$$

where $\boldsymbol{K}$ and $\boldsymbol{K}^i$ are the kernel matrices with respect to T and T^i, respectively. $\|\boldsymbol{K} - \boldsymbol{K}^i\|_2$ is the 2-norm of $[\boldsymbol{K} - \boldsymbol{K}^i]$, that is, the largest eigenvalue of $[\boldsymbol{K} - \boldsymbol{K}^i]$.

According to the above definition, the kernel stability is used to quantify the perturbation of the kernel matrix when an arbitrary example is removed. Different from the existing notions of stability, see, e.g., [31,16,5,21,29,12,34] and the references therein, our proposed stability is defined on the kernel matrix. Therefore, we can estimate its value from empirical data, which makes this stability usable for kernel selection in practice.

4.2 Upper Bounds via Kernel Stability

We will show that the kernel stability can yield the upper bounds of the variance of CV for KRR, LSSVM and SVM.

Kernel Ridge Regression. KRR has been successfully applied to solve regression problems, which is a special case of the regularized algorithms when the loss function

$$b = 0 \text{ and } \ell(f(\boldsymbol{x}), y) = (f(\boldsymbol{x}) - y)^2.$$

Theorem 1. *If the kernel function K is of β kernel stability, then for KRR,*

$$\operatorname*{var}_S(\ell_{f_{\text{kcv}}}(S)) \leq C_1\beta^2,$$

where $C_1 = 8\left(\frac{\kappa^2 M^2}{\lambda^3 m} + \frac{\kappa M^2}{\lambda^2 m}\right)^2$.

Proof. The proof is given in Appendix A.

This theorem shows that small β can restrict the value of $\operatorname{var}_S(\ell_{f_{\text{kcv}}}(S))$. Thus, we can select the kernel which has small β to prevent the over-fitting of CV caused by the high variance.

Least Squares Support Vector Machine. LSSVM is a popular learning machine for solving classification problems, its loss function is the square loss

$$\ell(f(\boldsymbol{x}) + b, y) = (y - f(\boldsymbol{x}) - b)^2.$$

Theorem 2. *If the kernel function K is of β kernel stability, then for LSSVM,*

$$\operatorname*{var}_S(\ell_{f_{\text{kcv}}}(S)) \leq C_2\beta^2,$$

where $C_2 = \left(\frac{2(\kappa+1)^2}{\lambda^3 m} + \frac{2(\kappa+1)}{\lambda^2 m}\right)^2$.

Proof. The proof is given in Appendix B.

Similar with KRR, this theorem also show that we can choose the kernel function which has small β to prevent the high variance for LSSVM.

Support Vector Machine. The loss function of SVM is the hinge loss

$$\ell(f(\boldsymbol{x}) + b, y) = \max\left(0, 1 - y\left(f(\boldsymbol{x}) + b\right)\right).$$

Theorem 3. *If the kernel function K is of β kernel stability, then for SVM,*

$$\operatorname*{var}_{S}(\ell_{f_{\mathrm{kcv}}}(S)) \leq C_3 \beta^{\frac{1}{2}} \left(1 + C_4 \beta^{\frac{1}{4}}\right)^2,$$

where $C_3 = 8\lambda^2\kappa^{\frac{3}{2}}$ *and* $C_4 = \left[\frac{1}{4\kappa}\right]^{\frac{1}{4}}$.

Proof. The proof is given in Appendix C.

The bound we obtain for SVM is different from our bounds for KRR and LSSVM. This is mainly due to the difference between the hinge loss and the squared loss.

5 Kernel Selection Criterion

Theorem 1, 2 and 3 show that the variance of CV can be bounded via the kernel stability. Thus, to prevent over-fitting caused by the high variance, it is reasonable to use the following criterion for kernel selection:

$$\arg\min_{K \in \mathcal{K}} \ell_{f_{\mathrm{kcv}}}(S) + \frac{\eta}{n}\beta,$$

where η is a trade-off parameter and $\mathcal{K}$ is an candidate set of kernel functions. However, by the definition of the kernel stability, we need to try all the possibilities of the training set to compute β, which is infeasible in practice. We should estimate it from the available empirical data. Therefore, we consider using the following kernel stability criterion in practice:

$$\arg\min_{K \in \mathcal{K}} k\text{-KS}(K) = \ell_{f_{\mathrm{kcv}}}(S) + \frac{\eta}{n} \cdot \max_{i \in \{1,\ldots,n\}} \|\boldsymbol{K} - \boldsymbol{K}^i\|_2.$$

This criterion consists of two parts: bias and variance. $\ell_{f_{\mathrm{kcv}}}(S)$ can be considered as the bias, and $\max_{i\in\{1,\ldots,n\}} \|\boldsymbol{K} - \boldsymbol{K}^i\|_2$ is the variance.

To apply this criterion, we should compute $\|\boldsymbol{K} - \boldsymbol{K}^i\|_2$, which requires the calculation of the eigenvalues of $[\boldsymbol{K} - \boldsymbol{K}^i]$, $i = 1, \ldots, n$. It is computationally expensive. Fortunately, this problem can be effectively solved by using the closed form of $\|\boldsymbol{K} - \boldsymbol{K}^i\|_2$ given by the following theorem.

Theorem 4. $\forall\ S \in \mathcal{Z}^n$ *and* $i \in \{1, \ldots, n\}$,

$$\|\boldsymbol{K} - \boldsymbol{K}^i\|_2 = \frac{\boldsymbol{K}_{ii} + \sqrt{\boldsymbol{K}_{ii}^2 + 4\sum_{j=1, j\neq i}^{n} \boldsymbol{K}_{ji}^2}}{2}.$$

Proof. By the definitions of $\boldsymbol{K}$ and $\boldsymbol{K}^i$, it is easy to verify that the characteristic polynomial of $[\boldsymbol{K} - \boldsymbol{K}^i]$ is

$$\det(t\boldsymbol{I} - (\boldsymbol{K} - \boldsymbol{K}^i)) = t^{n-2}(t^2 - \boldsymbol{K}_{ii}t - \sum_{j=1, j\neq i}^{n} \boldsymbol{K}_{ji}^2).$$

Thus, the eigenvalues of $\boldsymbol{K} - \boldsymbol{K}^i$ is

$$\sigma(\boldsymbol{K} - \boldsymbol{K}^i) = \left\{ \frac{\boldsymbol{K}_{ii} \pm \sqrt{\boldsymbol{K}_{ii}^2 + 4\sum_{j=1, j\neq i}^{n} \boldsymbol{K}_{ji}^2}}{2}, \overbrace{0, \ldots, 0}^{n-2} \right\}.$$

So, the largest eigenvalue is

$$\frac{\boldsymbol{K}_{ii} + \sqrt{\boldsymbol{K}_{ii}^2 + 4\sum_{j=1, j\neq i}^{n} \boldsymbol{K}_{ji}^2}}{2}.$$

Hence we complete the proof of Theorem 4.

This theorem shows that only $O(n^2)$ is needed to compute

$$\max_{i\in\{1,\ldots,n\}} \|\boldsymbol{K} - \boldsymbol{K}^i\|_2,$$

making the criterion based on kernel stability computationally efficient.

Remark 1. Instead of choosing a single kernel, several authors consider combining multiple kernels by some criteria, called multiple kernel learning (MKL), see, e.g., [22,1,30,18,25], etc. Our criterion k-KS(K) can also be applied to MKL:

$$\underset{\boldsymbol{\mu}=(\mu_1,\ldots,\mu_k)}{\arg\min} \; k\text{-KS}(K_{\boldsymbol{\mu}}), \text{s.t.} \|\boldsymbol{\mu}\|_p = 1, \boldsymbol{\mu} \geq 0,$$

where $K_{\boldsymbol{\mu}} = \sum_{i=1}^{k} \mu_i K_i$, which can be efficiently solved using gradient-based algorithms [17]. However, in this paper we mainly want to verify the effectiveness of our kernel stability criterion. Therefore, in our experiments, we focus on comparing our criterion with other popular kernel selection criteria.

5.1 Time Complexity Analysis

To compute our kernel stability criterion k-KS(K), we need kF to calculate $\ell_{f_{k\text{cv}}}(S)$, where F is the time complexity of training on the data set of size $(k-1)k/n$, n is the size of the training set. We also need $O(n^2)$ to compute

$$\max_{i\in\{1,\ldots,n\}} \|\boldsymbol{K} - \boldsymbol{K}^i\|_2.$$

Thus, the overall time complexity of k-KS(K) is

$$O(kF + n^2).$$

Remark 2. In our previous work [24], we presented a strategy for approximating the k-fold CV based on the Bouligand influence function [10]. This approximate method requires the solution of the algorithm only once, which can dramatically improve the efficiency. Thus, the time complexity of the approximate k-KS(K) can reduce to $O(F + n^2)$.

6 Experiments

In this section, we will compare our proposed kernel selection criteria (k-KS, $k = 5, 10$) with 5 popular kernel selection criteria: 5-fold cross-validation (5-CV), 10-fold cross-validation (10-CV), the efficient leave-one-out cross-validation (ELOO) [6], Bayesian regularisation (BR) [7], and the latest eigenvalues perturbation criterion (EP) [23]. The evaluation is made on 9 popular data sets from LIBSVM Data. All data sets are normalized to have zero-means and unit-variances on every attribute to avoid numerical problems. We use the popular Gaussian kernels

$$K(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|_2^2}{2\tau}\right)$$

as our candidate kernels,

$$\tau \in \{2^i, i = -10, -9, \ldots, 10\}.$$

For each data set, we have run all the methods 10 times with randomly selected 70% of all data for training and the other 30% for test. The learning machine we considered is LSSVM.

6.1 Accuracy

In this subsection, we will compare the performance of 5-KS (ours), 10-KS (ours), 5-CV, 10-CV, ELOO, BR and EP. In the first experiment, we set $\eta = 1$ (the parameter of the 5-KS and 10-KS criterion, we will explore the effect of this parameter in the next experiment). The average test errors are reported in Table 1. The elements in this table are obtained as follows: For each training set and each regularization parameter[1] λ, $\lambda \in \{10^i, i = -4, \ldots, -1\}$, we choose the kernel by each kernel selection criterion on the training set, and evaluate the test error for the chosen parameters on the test set. The results in Table 1 can be summarized as follows: (a) k-KS gives better results than k-CV on most data sets, $k = 5, 10$. In particular, for each λ, k-KS outperforms k-CV on 8 (or more) out of 9 sets, and also give results closed to results of k-CV on the remaining set. Thus, it indicates that using the kernel stability to restrict the high variance

[1] the value of λ we set seems too small at first sight, but in fact, the regularized algorithm we considered in this paper is $\frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i), y_i) + \lambda\|f\|_K^2$, while other authors usually ignore $1/n$. Therefore, the value of λ in our paper is $1/n$ time of that of regularized algorithms other authors considered.

Table 1. The test errors (%) with standard deviations of 5-KS (ours), 10-KS (ours), 5-CV,10-CV, ELOO, BR and EP. For each training set, each regularization parameter λ ($\lambda \in \{10^{-i}, i = -4, \ldots, -1\}$), we choose the kernel by each kernel selection criterion on the training set, and evaluate the test error for the chosen kernel on test set.

$\lambda = 0.0001$							
Method	5-CV	5-KS	10-CV	10-KS	ELOO	BR	EP
australian	14.78 ± 2.4	**13.23** ± 2.0	15.46 ± 1.6	15.27 ± 1.4	14.30 ± 2.6	14.30 ± 2.6	14.15 ± 2.9
heart	19.01 ± 3.9	18.52 ± 3.1	18.27 ± 3.7	**18.17** ± 2.3	18.52 ± 3.4	18.52 ± 3.3	17.83 ± 6.1
ionosphere	4.76 ± 1.5	4.38 ± 1.5	4.67 ± 1.7	**4.35** ± 1.9	5.33 ± 2.2	5.13 ± 2.1	6.76 ± 3.9
breast	3.61 ± 0.8	**3.41** ± 0.7	3.51 ± 0.6	3.45 ± 0.6	3.41 ± 0.9	3.41 ± 0.9	5.27 ± 1.4
diabetes	23.65 ± 3.7	23.65 ± 3.7	23.83 ± 3.3	23.91 ± 2.9	23.04 ± 2.5	**22.96** ± 2.7	30.26 ± 2.3
german	24.60 ± 1.3	**24.17** ± 1.2	25.80 ± 1.1	24.67 ± 1.3	24.67 ± 1.4	24.60 ± 1.3	29.67 ± 3.1
liver	26.92 ± 2.5	27.12 ± 2.0	27.50 ± 2.6	**26.15** ± 1.0	26.55 ± 1.3	26.73 ± 2.0	30.08 ± 4.7
sonar	14.19 ± 4.0	13.23 ± 3.5	13.87 ± 2.9	**12.58** ± 2.4	12.90 ± 4.9	12.90 ± 4.9	17.74 ± 6.7
a2a	18.44 ± 1.0	17.14 ± 0.9	17.94 ± 1.0	**15.20** ± 0.6	17.91 ± 1.0	17.34 ± 1.0	18.92 ± 1.5
$\lambda = 0.001$							
Method	5-CV	5-KS	10-CV	10-KS	ELOO	BR	EP
australian	14.30 ± 1.2	**12.19** ± 1.3	13.30 ± 0.7	12.30 ± 0.4	14.43 ± 0.9	13.43 ± 0.9	16.29 ± 3.4
heart	18.27 ± 6.7	15.80 ± 4.0	17.80 ± 4.3	15.31 ± 4.2	14.83 ± 4.8	**14.07** ± 4.9	20.77 ± 6.3
ionosphere	5.33 ± 1.9	**3.81** ± 1.9	5.33 ± 1.9	4.38 ± 1.0	6.48 ± 2.2	6.48 ± 2.2	5.38 ± 4.6
breast	3.61 ± 0.8	3.42 ± 0.8	3.51 ± 0.8	**3.22** ± 0.6	3.32 ± 0.8	3.32 ± 0.8	5.56 ± 1.1
diabetes	23.65 ± 2.4	23.48 ± 1.8	23.83 ± 2.3	23.45 ± 2.2	24.22 ± 1.9	**23.22** ± 1.9	26.52 ± 0.4
german	25.07 ± 2.4	**24.60** ± 2.4	23.93 ± 1.1	23.87 ± 0.9	24.60 ± 2.4	24.67 ± 2.4	25.13 ± 2.1
liver	29.04 ± 3.5	28.46 ± 3.3	27.12 ± 4.6	**26.54** ± 1.8	27.12 ± 2.7	26.82 ± 2.6	28.46 ± 3.3
sonar	14.84 ± 6.7	13.87 ± 4.5	11.61 ± 5.9	**11.55** ± 5.7	13.55 ± 6.7	13.83 ± 6.8	13.90 ± 7.3
a2a	17.23 ± 1.1	**15.51** ± 0.9	17.35 ± 1.0	16.11 ± 1.0	16.91 ± 0.9	16.94 ± 0.9	19.71 ± 1.1
$\lambda = 0.01$							
Method	5-CV	5-KS	10-CV	10-KS	ELOO	BR	EP
australian	14.59 ± 2.0	13.82 ± 1.9	14.98 ± 2.0	**13.72** ± 2.0	14.01 ± 2.1	14.01 ± 2.1	16.54 ± 3.4
heart	18.27 ± 2.3	17.78 ± 2.2	18.27 ± 1.6	18.02 ± 0.6	**17.28** ± 1.5	17.78 ± 2.0	19.74 ± 6.6
ionosphere	4.95 ± 1.5	**4.38** ± 1.2	4.95 ± 1.5	5.14 ± 1.2	5.14 ± 2.1	5.14 ± 2.1	9.52 ± 3.3
breast	3.51 ± 0.7	3.46 ± 0.6	3.80 ± 0.6	3.80 ± 0.6	3.75 ± 1.3	**3.41** ± 1.0	7.02 ± 0.8
diabetes	24.00 ± 1.4	**22.30** ± 1.3	23.48 ± 1.4	23.83 ± 1.3	23.83 ± 1.7	23.83 ± 1.7	25.83 ± 1.8
german	26.40 ± 0.9	26.33 ± 0.7	26.47 ± 0.9	**24.93** ± 0.4	26.87 ± 1.1	26.25 ± 1.5	28.67 ± 1.3
liver	28.85 ± 1.8	**25.27** ± 1.2	30.00 ± 2.6	28.65 ± 2.4	28.46 ± 2.3	28.65 ± 2.3	29.42 ± 3.0
sonar	14.52 ± 5.4	13.55 ± 2.9	13.23 ± 4.4	12.23 ± 3.5	12.58 ± 3.8	**11.94** ± 4.0	14.74 ± 4.0
a2a	18.88 ± 2.0	**17.76** ± 1.1	18.97 ± 1.9	17.82 ± 1.7	18.76 ± 2.1	18.41 ± 2.5	20.15 ± 2.5
$\lambda = 0.1$							
Method	5-CV	5-KS	10-CV	10-KS	ELOO	BR	EP
australian	14.30 ± 2.1	13.53 ± 1.8	14.30 ± 2.1	14.20 ± 0.7	13.91 ± 1.4	**13.41** ± 1.7	14.54 ± 2.9
heart	19.51 ± 3.2	19.26 ± 2.9	19.26 ± 3.4	**18.26** ± 2.9	19.51 ± 3.2	19.26 ± 3.4	22.65 ± 4.0
ionosphere	12.76 ± 10.5	8.38 ± 3.7	9.14 ± 4.2	**8.28** ± 3.5	9.33 ± 4.3	9.14 ± 4.6	12.95 ± 3.9
breast	3.92 ± 1.4	3.22 ± 1.3	3.62 ± 1.3	**3.12** ± 1.4	3.41 ± 1.1	3.21 ± 1.6	4.98 ± 1.0
diabetes	29.65 ± 1.8	29.83 ± 2.0	29.74 ± 2.1	**29.48** ± 1.8	29.57 ± 1.7	29.57 ± 1.5	35.65 ± 1.4
german	31.21 ± 1.7	27.40 ± 1.5	27.40 ± 1.4	26.51 ± 1.2	**25.40** ± 1.1	29.40 ± 1.6	31.40 ± 1.4
liver	33.46 ± 7.4	**31.08** ± 7.0	33.08 ± 8.0	32.42 ± 6.3	31.85 ± 8.6	38.65 ± 5.8	33.08 ± 8.0
sonar	27.81 ± 9.6	**26.06** ± 9.3	27.42 ± 9.3	27.10 ± 8.7	26.77 ± 9.3	26.77 ± 9.3	27.10 ± 5.0
a2a	24.68 ± 1.7	**22.18** ± 1.5	25.68 ± 1.9	22.68 ± 1.8	24.68 ± 1.7	24.31 ± 0.8	23.21 ± 1.7

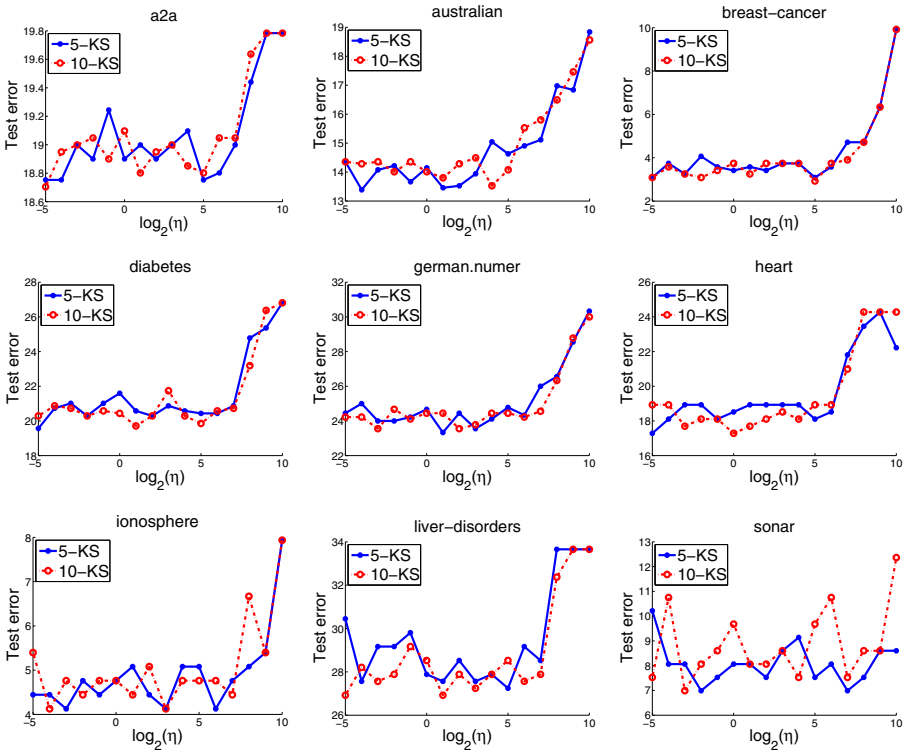

Fig. 1. The average test errors using 5-KS and 10-KS on different η. The regularization parameter λ is set as 0.001 (in Table 1, one can see that for most data sets, $\lambda = 0.001$ can achieve good results. Thus, we only consider setting $\lambda = 0.001$). For each training set, each η, we choose the kernel by 5-KS and 10-KS kernel selection criteria on the training set, and evaluate the test errors for the chosen parameters on test set.

of cross-validation can guarantee good generalization performance; (b) k-KS is better than BR on most data sets. In particular, for each λ, k-KS outperforms k-CV on 6 (or more) out of 9 sets; (c) BR is comparable or better than ELOO on most data sets; (d) The performances of the 5-KS and 10-KS are comparable.

6.2 Effect of the Parameter η

In this experiments, we will explore the effect of the η. The average test errors on different η are given in Figure 1. For each training set, each η, we choose the kernel by 5-KS and 10-KS kernel selection criteria on the training set, and evaluate the test errors for the chosen parameters on test set. It turns out that η is robust, and the test errors are not very sensitive w.r.t $\eta \in [2^{-2}, 2^5]$ on most data sets. Moreover, we find that $\eta \in [2^{-2}, 2^5]$ is a good choice for k-KS. Thus, we can select $\eta \in [2^{-2}, 2^5]$ in practice.

7 Conclusion

We propose a novel kernel selection criterion via a newly defined concept of kernel stability, which can prevent over-fitting of cross-validation (CV) caused by high variance. We illuminate that the variance of CV for KRR, LSSVM and SVM can be bounded with the kernel stability, so we can use this stability to control the variance of CV to avoid over-fitting. Moreover, we derive a closed form of the estimate of the kernel stability, making the kernel selection criterion based on the kernel stability computationally efficient and practically useful. Finally, our kernel selection criterion is theoretically justified and experimentally validated. To our knowledge, this is the first attempt to use the notion of stability to control the variance of CV for kernel selection in kernel methods.

Future work includes extending our method to other kernel based methods and multiple kernel learning, and using the notion of the kernel stability to derive the generalization error bounds for kernel methods.

Acknowledgments. The work is supported in part by the National Natural Science Foundation of China under grant No. 61170019.

Appendix A: Proof of Theorem 1

Lemma 1 (Proposition 1 in [13]). *Let h' denote the hypothesis returned by KRR when using the approximate kernel matrix $\boldsymbol{K}'$. Then, the following inequality holds for all $\boldsymbol{x} \in \mathcal{X}$:*

$$|h'(\boldsymbol{x}) - h(\boldsymbol{x})| \leq \frac{\kappa M}{\lambda^2 m} \|\boldsymbol{K}' - \boldsymbol{K}\|_2.$$

Definition 2 (Loss stability [20]). *The loss stability of a learning algorithm A trained on m examples and with respect to a loss ℓ is defined as*

$$\mathrm{ls}_{m,\ell}(A) = \mathbb{E}_{T:|T|=m,z',z}\left[\left(\ell'_{A(T)}(z) - \ell'_{A(T^{z'})}(z)\right)^2\right],$$

where $T^{z'}$ denote the set of examples obtained by replacing an example chosen uniformly at random from T by z'. A learning algorithm A is γ-loss stable if $\mathrm{ls}_{m,\ell}(A) \leq \gamma$.

Lemma 2 (Theorem 1 in [20]). *Consider any learning algorithm A that is γ-loss stable with respect to ℓ. Then*

$$\operatorname*{var}_S(\ell_{f_{k\mathrm{cv}}}(S)) \leq \frac{1}{k} \operatorname*{var}_S(\ell_{f_{S\setminus S_1}}(S_1)) + \left(1 - \frac{1}{k}\right)\gamma.$$

Proof (of Theorem 1). Note that $f_{T^i}(\boldsymbol{x})$ is the hypothesis returned by KRR using $\boldsymbol{K}^i$. According to the definition of β kernel stability, we have $\|\boldsymbol{K} - \boldsymbol{K}^i\|_2 \leq \beta$. By Lemma 1,

$$|f_T(\boldsymbol{x}) - f_{T^i}(\boldsymbol{x})| \leq \frac{\kappa M}{\lambda^2 m} \|\boldsymbol{K}^i - \boldsymbol{K}\|_2 \leq \frac{\beta \kappa M}{\lambda^2 m}. \tag{1}$$

Since $f_T(\boldsymbol{x}) = \sum_{i=1}^{m} \alpha_i K(\boldsymbol{x}, \boldsymbol{x}_i) = \boldsymbol{k}_{\boldsymbol{x}}\boldsymbol{\alpha}$, where $\boldsymbol{\alpha} = [\boldsymbol{K} + m\lambda\boldsymbol{I}]^{-1}\boldsymbol{y}$ and $\boldsymbol{k}_{\boldsymbol{x}} = (K(\boldsymbol{x}, \boldsymbol{x}_1), \ldots, K(\boldsymbol{x}, \boldsymbol{x}_m))^{\mathrm{T}}$. Thus, we have

$$\begin{aligned} |f_T(\boldsymbol{x})| &= |\boldsymbol{k}_{\boldsymbol{x}}^{\mathrm{T}}\boldsymbol{\alpha}| = |\boldsymbol{k}_{\boldsymbol{x}}[\boldsymbol{K} + m\lambda\boldsymbol{I}]^{-1}\boldsymbol{y}| \\ &\leq \|\boldsymbol{k}_{\boldsymbol{x}}\|\|\boldsymbol{y}\|\|[\boldsymbol{K} + m\lambda\boldsymbol{I}]^{-1}\|_2 \\ &\leq \frac{\kappa\sqrt{m}M\sqrt{m}}{m\lambda} \\ &= \frac{\kappa M}{\lambda}. \end{aligned} \tag{2}$$

Thus, $\forall \boldsymbol{z} \in \mathcal{Z}$, $\forall\, T \in \mathcal{Z}^m$ and $\forall i \in \{1, \ldots, m\}$,

$$\begin{aligned} &|\ell_{f_T}(z) - \ell_{f_{T^i}}(z)| \\ &= |(f_T(\boldsymbol{x}) - y)^2 - (f_{T^i}(\boldsymbol{x}) - y)^2| \\ &= |(f_T(\boldsymbol{x}) - f_{T^i}(\boldsymbol{x}))(f_T(\boldsymbol{x}) + f_{T^i}(\boldsymbol{x}) - 2y)| \\ &\leq \left(\frac{\beta\kappa M}{\lambda^2 m}\right) \cdot \left(\frac{2\kappa M}{\lambda} + 2M\right). \end{aligned} \tag{3}$$

According to Lemma 2 in [20], we have

$$\mathrm{ls}_{m,\ell}(A) \leq \mathbb{E}_{T,z',z}\left[\left(\ell_{A(T)}(z) - \ell_{A(T^{z'})}(z)\right)^2\right]. \tag{4}$$

So, from (3), $\forall T, \forall z, \forall z'$, $\left(\ell_{f_T}(z) - \ell_{f_{T^{z'}}}(z)\right)^2 \leq$

$$\begin{aligned} &\leq \left(\left|\ell_{f_T}(z) - \ell_{f_{T^i}}(z)\right| + \left|\ell_{f_{T^i}}(z) - \ell_{A(T^{z'})}(z)\right|\right)^2 \\ &\leq \left(2\beta\left(\frac{2\kappa^2 M}{\lambda^3 m} + \frac{2\kappa M^2}{\lambda^2 m}\right)\right)^2 \\ &= C_1\beta^2. \end{aligned}$$

Thus, according to (4), we have

$$\mathrm{ls}_{m,\ell}(A) \leq C_1\beta^2 = \gamma. \tag{5}$$

According to Lemma 5 in [20], we have

$$\begin{aligned} &\operatorname*{var}_S(\ell_{f_{S\setminus S_1}}(S_1)) = \mathrm{cov}(\ell_{f_{S\setminus S_1}}(S_1), \ell_{f_{S\setminus S_1}}(S_1)) \\ &= \mathbb{E}_{S\setminus S_1, z_1', z_2}\left[\left(\ell'_{f_{S\setminus S_1}}(z_2) - \ell'_{f_{(S\setminus S_1)^{z_1'}}}(z_2)\right)^2\right] \\ &= \mathrm{ls}_{m,\ell}(A) \\ &\leq C_1\beta^2 = \gamma \text{ (According to (5)).} \end{aligned} \tag{6}$$

Substituting (5) and (6) into Lemma 2, we complete the proof of Theorem 1.

Appendix B: Proof of Theorem 2

Proof (of Theorem 2). For LSSVM,

$$f_T(\boldsymbol{x}) = \boldsymbol{k}_{\boldsymbol{x}}^{\mathrm{T}}\boldsymbol{\alpha} + b = \tilde{\boldsymbol{k}}_{\boldsymbol{x}}\boldsymbol{M}^{-1}\tilde{\boldsymbol{y}},$$

where

$$\tilde{\boldsymbol{k}}_{\boldsymbol{x}} = (K(\boldsymbol{x}, \boldsymbol{x}_1), \ldots, K(\boldsymbol{x}, \boldsymbol{x}_m), 1), \tilde{\boldsymbol{y}} = [y_1, \ldots, y_m, 0]^{\mathrm{T}}$$

and

$$\boldsymbol{M} = \begin{bmatrix} \boldsymbol{K} + m\lambda\boldsymbol{I} & \mathbf{1} \\ \mathbf{1}^{\mathrm{T}} & 0 \end{bmatrix},$$

Thus, it is easy to verify that

$$|f_T(\boldsymbol{x}) - f_{T^i}(\boldsymbol{x})| = |\tilde{\boldsymbol{k}}_{\boldsymbol{x}}(\boldsymbol{M}^{-1}\tilde{\boldsymbol{y}} - \boldsymbol{M}_i^{-1}\tilde{\boldsymbol{y}})|,$$

where

$$\boldsymbol{M}_i = \begin{bmatrix} \boldsymbol{K}^i + m\lambda\boldsymbol{I} & \mathbf{1} \\ \mathbf{1}^{\mathrm{T}} & 0 \end{bmatrix}.$$

Thus,

$$\begin{aligned} |f_T(\boldsymbol{x}) - f_{T^i}(\boldsymbol{x})| &\leq \|\tilde{\boldsymbol{k}}_{\boldsymbol{x}}\|\|\boldsymbol{M}^{-1} - \boldsymbol{M}_i^{-1}\|_2\|\tilde{\boldsymbol{y}}\| \\ &\leq \sqrt{m\kappa^2 + 1}\|\boldsymbol{M}^{-1} - \boldsymbol{M}_i^{-1}\|\sqrt{m} \\ &\leq m(\kappa+1)\|\boldsymbol{M}^{-1}(\boldsymbol{M} - \boldsymbol{M}_i)\boldsymbol{M}_i^{-1}\|_2 \\ &\leq m(\kappa+1)\|\boldsymbol{M}^{-1}\|_2\|\boldsymbol{M} - \boldsymbol{M}_i\|_2\|\boldsymbol{M}_i^{-1}\|_2 \\ &\leq m(\kappa+1)\frac{\|\boldsymbol{M} - \boldsymbol{M}_i\|_2}{m^2\lambda^2} \\ &\leq \frac{\kappa+1}{m\lambda^2}\beta. \end{aligned}$$

Similar with the proof of Eq (2), we can obtain $f_T(\boldsymbol{x}) \leq \frac{\kappa+1}{\lambda}$. Thus, we have

$$\begin{aligned} &|\ell_{f_T}(z) - \ell_{f_{T^i}}(z)| = |(f_T(\boldsymbol{x}) - y)^2 - (f_{T^i}(\boldsymbol{x}) - y)^2| \\ &= |(f_T(\boldsymbol{x}) - f_{T^i}(\boldsymbol{x}))(f_T(\boldsymbol{x}) + f_{T^i}(\boldsymbol{x}) - 2y)| \\ &\leq \left(\frac{\kappa+1}{m\lambda^2}\beta\right)\left(\frac{2\kappa+2}{\lambda} + 2\right). \end{aligned}$$

Similar with the proof of (5) and (6), it is easy to verify that

$$\mathrm{ls}_{m,\ell}(A) \leq C_2\beta^2 = \gamma$$

and

$$\operatorname*{var}_S(\ell_{f_{S\setminus S_1}}(S_1)) \leq C_2\beta^2 = \gamma.$$

From Lemma 2, we prove Theorem 2.

Appendix C: Proof of Theorem3

Lemma 3 (Proposition 2 in [13]). *Let h' denote the hypothesis returned by SVMs when using the approximate kernel matrix $\boldsymbol{K}'$. Then, the following inequality holds for all $\boldsymbol{x} \in \mathcal{X}$:*

$$|h'(\boldsymbol{x}) - h(\boldsymbol{x})| \leq \sqrt{2}\lambda\kappa^{\frac{3}{4}}\|\boldsymbol{K}' - \boldsymbol{K}\|_2^{\frac{1}{4}}\left[1 + \left[\frac{\|\boldsymbol{K}' - \boldsymbol{K}\|_2}{4\kappa}\right]^{\frac{1}{4}}\right].$$

Proof (of Theorem 3). Note that $f_{T^i}(\boldsymbol{x})$ is the hypothesis returned by SVM using the ith removed kernel matrix $\boldsymbol{K}^i$. By Lemma 3 and the definition of β kernel stability,

$$|f_T(\boldsymbol{x}) - f_{T^i}(\boldsymbol{x})| \leq \sqrt{2}\lambda\kappa^{\frac{3}{4}}\beta^{\frac{1}{4}}\left[1 + \left[\frac{\beta}{4\kappa}\right]^{\frac{1}{4}}\right].$$

Since the hinge loss ℓ is 1-Lipschitz, so $\forall \boldsymbol{z}, T, z'$

$$|\ell_{f_T}(z) - \ell_{f_{T^i}}(z)| \leq \sqrt{2}\lambda\kappa^{\frac{3}{4}}\beta^{\frac{1}{4}}\left[1 + \left[\frac{\beta}{4\kappa}\right]^{\frac{1}{4}}\right].$$

Similar with the proof of (5) and (6), we can obtain that

$$\mathrm{ls}_{m,\ell}(A) \leq C_3\beta^{\frac{1}{2}}\left(1 + C_3\beta^{\frac{1}{4}}\right)^2 = \gamma$$

and

$$\underset{S}{\mathrm{var}}(\ell_{f_{S\setminus S_1}}(S_1)) \leq C_3\beta^{\frac{1}{2}}\left(1 + C_3\beta^{\frac{1}{4}}\right)^2 = \gamma.$$

Thus, Theorem 3 follows from substituting the above two equations to Lemma 2.

References

1. Bach, F., Lanckriet, G., Jordan, M.: Multiple kernel learning, conic duality, and the SMO algorithm. In: Proceedings of the 21st International Conference on Machine Learning (ICML 2004), pp. 41–48 (2004)
2. Bartlett, P.L., Mendelson, S.: Rademacher and Gaussian complexities: Risk bounds and structural results. Journal of Machine Learning Research 3, 463–482 (2002)
3. Bengio, Y., Grandvalet, Y.: No unbiased estimator of the variance of k-fold cross-validation. Journal of Machine Learning Research 5, 1089–1105 (2004)
4. Blum, A., Kalai, A., Langford, J.: Beating the hold-out: Bounds for k-fold and progressive cross-validation. In: Proceedings of the 12nd Annual Conference on Computational Learning Theory (COLT 1999), pp. 203–208 (1999)
5. Bousquet, O., Elisseeff, A.: Stability and generalization. Journal of Machine Learning Research 2, 499–526 (2002)
6. Cawley, G.C.: Leave-one-out cross-validation based model selection criteria for weighted LS-SVMs. In: Proceeding of the International Joint Conference on Neural Networks (IJCNN 2006), pp. 1661–1668 (2006)

7. Cawley, G.C., Talbot, N.L.C.: Preventing over-fitting during model selection via Bayesian regularisation of the hyper-parameters. Journal of Machine Learning Research 8, 841–861 (2007)
8. Cawley, G.C., Talbot, N.L.C.: On over-fitting in model selection and subsequent selection bias in performance evaluation. Journal of Machine Learning Research 11, 2079–2107 (2010)
9. Chapelle, O., Vapnik, V., Bousquet, O., Mukherjee, S.: Choosing multiple parameters for support vector machines. Machine Learning 46(1-3), 131–159 (2002)
10. Christmann, A., Messem, A.V.: Bouligand derivatives and robustness of support vector machines for regression. Journal of Machine Learning Research 9, 915–936 (2008)
11. Cortes, C., Kloft, M., Mohri, M.: Learning kernels using local Rademacher complexity. In: Advances in Neural Information Processing Systems 25 (NIPS 2013), pp. 2760–2768. MIT Press (2013)
12. Cortes, C., Mohri, M., Pechyony, D., Rastogi, A.: Stability of transductive regression algorithms. In: Proceedings of the 25th International Conference on Machine Learning (ICML 2008), pp. 176–183 (2008)
13. Cortes, C., Mohri, M., Talwalkar, A.: On the impact of kernel approximation on learning accuracy. In: Proceedings of the 13rd International Conference on Artificial Intelligence and Statistics (AISTATS 2010), pp. 113–120 (2010)
14. Debruyne, M., Hubert, M., Suykens, J.A.: Model selection in kernel based regression using the influence function. Journal of Machine Learning Research 9, 2377–2400 (2008)
15. Geras, K.J., Sutton, C.: Multiple-source cross-validation. In: Proceedings of the 30th International Conference on Machine Learning (ICML 2013), pp. 1292–1300 (2013)
16. Kearns, M.J., Ron, D.: Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. Neural Computation 11(6), 1427–1453 (1999)
17. Keerthi, S.S., Sindhwani, V., Chapelle, O.: An efficient method for gradient-based adaptation of hyperparameters in SVM models. In: Advances in Neural Information Processing Systems 19 (NIPS 2007), pp. 673–680. MIT Press (2007)
18. Kloft, M., Brefeld, U., Sonnenburg, S., Zien, A.: l_p-norm multiple kernel learning. Journal of Machine Learning Research 12, 953–997 (2011)
19. Kohavi, R.: A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the 14th International Conference on Artificial Intelligence (IJCAI 1995), pp. 1137–1143 (1995)
20. Kumar, R., Lokshtanov, D., Vassilvitskii, S., Vattani, A.: Near-optimal bounds for cross-validation via loss stability. In: Proceedings of the 30th International Conference on Machine Learning (ICML 2013), pp. 27–35 (2013)
21. Kutin, S., Niyogi, P.: Almost-everywhere algorithmic stability and generalization error. In: Proceedings of the 18th Conference in Uncertainty in Artificial Intelligence (UAI 2002), pp. 275–282 (2002)
22. Lanckriet, G.R.G., Cristianini, N., Bartlett, P.L., Ghaoui, L.E., Jordan, M.I.: Learning the kernel matrix with semidefinite programming. Journal of Machine Learning Research 5, 27–72 (2004)
23. Liu, Y., Jiang, S., Liao, S.: Eigenvalues perturbation of integral operator for kernel selection. In: Proceedings of the 22nd ACM International Conference on Information and Knowledge Management (CIKM 2013), pp. 2189–2198 (2013)
24. Liu, Y., Jiang, S., Liao, S.: Efficient approximation of cross-validation for kernel methods using Bouligand influence function. In: Proceedings of the 31st International Conference on Machine Learning (ICML 2014 (1)), pp. 324–332 (2014)

25. Liu, Y., Liao, S., Hou, Y.: Learning kernels with upper bounds of leave-one-out error. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM 2011), pp. 2205–2208 (2011)
26. Luxburg, U.V., Bousquet, O., Schölkopf, B.: A compression approach to support vector model selection. Journal of Machine Learning Research 5, 293–323 (2004)
27. Ng, A.Y.: Preventing "overfitting" of cross-validation data. In: Proceeding of the 14th International Conference on Machine Learning (ICML 1997), pp. 245–253 (1997)
28. Nguyen, C.H., Ho, T.B.: Kernel matrix evaluation. In: Proceedings of the 20th International Joint Conference on Artifficial Intelligence (IJCAI 2007), pp. 987–992 (2007)
29. Poggio, T., Rifkin, R.M., Mukherjee, S., Niyogi, P.: General conditions for predictivity in learning theory. Nature 428(6981), 419–422 (2004)
30. Rakotomamonjy, A., Bach, F., Canu, S., Grandvalet, Y.: SimpleMKL. Journal of Machine Learning Research 9, 2491–2521 (2008)
31. Rogers, W.H., Wagner, T.J.: A finite sample distribution-free performance bound for local discrimination rules. The Annals of Statistics 6, 506–514 (1978)
32. Saunders, C., Gammerman, A., Vovk, V.: Ridge regression learning algorithm in dual variables. In: Proceedings of the 15th International Conference on Machine Learning (ICML 1998), pp. 515–521 (1998)
33. Schölkopf, B., Smola, A.J.: Learning with kernels. MIT Press, Cambridge (2002)
34. Shalev-Shwartz, S., Shamir, O., Srebro, N., Sridharan, K.: Learnability, stability and uniform convergence. Journal of Machine Learning Research 11, 2635–2670 (2010)
35. Suykens, J.A.K., Vandewalle, J.: Least squares support vector machine classifiers. Neural Processing Letters 9(3), 293–300 (1999)
36. Vapnik, V.: The nature of statistical learning theory. Springer (2000)